Zootaxa 2946: 45–51 (2011)
www.mapress.com/zootaxa/

Editorial

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

Exploring character conflict in molecular data*

ROBERT H. CRUICKSHANK
Department of Ecology, Faculty of Agriculture and Life Sciences, PO Box 84, Lincoln University, Lincoln 7647, Christchurch, New Zealand. E-mail: Rob.Cruickshank@lincoln.ac.nz

**In*: Carvalho, M.R. de & Craig, M.T. (Eds) (2011) Morphological and Molecular Approaches to the Phylogeny of Fishes: Integration or Conflict?. *Zootaxa*, 2946, 1–142.

Abstract

Mooi & Gill (2010) have made a number of criticisms of statistical approaches to the phylogenetic analysis of molecular data as it is currently practiced. There are many different uses for molecular phylogenies, and for most of them statistical methods are entirely appropriate, but for taxonomic purposes the way that these methods have been used is questionable. In these cases it is necessary to introduce an extra step into the analysis – exploration of character conflict. Existing methods for exploring character conflict in molecular data such as spectral analysis, phylogenetic networks, likelihood mapping and sliding window analyses are briefly reviewed, but there is also a need for development of new tools to facilitate the analysis of large data sets. Incorporation of previous phylogenies as priors in Bayesian analyses could help to provide taxonomic stability, while still leaving room for new data to alter these conclusions if they contain sufficiently strong phylogenetic signal. Molecular phylogeneticists should make a clearer distinction between the different uses to which their phylogenies are put; methods suitable in one context may not be appropriate in others.

Key words: molecular systematics, character conflict, spectral analysis, phylogenetic networks, likelihood mapping, sliding window analyses, Bayesian priors, taxonomic stability

Introduction

In a recent issue of *Zootaxa*, Mooi and Gill (2010) criticise the way that molecular data are typically used in taxonomy. They raise a number of interesting and important points and make four recommendations that they believe would bring molecular systematics "back to its fundamental principles" (p.26). The first of these is that molecular taxonomists should "examine data quality, character distribution, and evidence; plot characters to identify and examine character conflict, and weigh evidence for homology" (p.26). Identification and exploration of character conflict does indeed appear to be a missing step in many, if not most, molecular phylogenetic studies. There are a number of possible reasons for this, some of which are discussed below. The purpose of this brief review is to draw attention to several tools for exploring character conflict in molecular data that are freely available but not widely used, in the hope that more widespread adoption of these techniques may begin to satisfy at least some of the criticisms of Mooi & Gill.

Background

There are many different uses for molecular phylogenies and reasons for constructing them. For most of these purposes a statistical approach is entirely appropriate. For example, evolutionary biologists interested reconstructing ancestral character states, or ecologists wanting to examine the phylogenetic structure of ecological communities, can use Bayesian approaches to incorporate phylogenetic uncertainty into their analyses (Huelsenbeck et al 2001).

Accepted by M.R. de Carvalho: 2 May 2011; published: 8 Jul. 2011

Another important use of molecular phylogenies is taxonomy. It is not really possible to incorporate phylogenetic uncertainty into classifications, and therefore a decision must be made as to the best estimate of the phylogenetic relationships of the taxa in question. Some of the uncertainty in our phylogenies is due to lack of signal, but some is due to conflict between different signals. Statistical approaches allow all conflicting signals to be considered in downstream applications, weighted according to the evidence in favour of each of them, but for taxonomic purposes it is necessary to resolve the conflict and decide which of the conflicting signals is most likely to reflect the actual evolutionary history of the group. This assessment will, of course, always be provisional, as new data may cause us to change our opinion, which is why classifications change over time, but as we accumulate data it is expected that we will converge on a stable and lasting taxonomy based on the evolutionary relationships of the organisms in our classification. This should be the goal of all taxonomists.

Recently, there has been a great deal of excitement among molecular phylogeneticists about new statistical methods for phylogenetics, including Bayesian, information theoretic and coalescent-based model-fitting approaches (e.g. Edwards 2009). These have been accompanied by a corresponding increase in computing power and more efficient algorithms for searching tree space that have meant that these computationally intensive methods are now possible for real data sets (although there has also been a general increase in the size of data sets that has, to some extent, offset this). For many applications these powerful new statistical methods allow analyses that have not previously been possible, and for this reason molecular phylogeneticists have, quite rightly, been quick to adopt and develop them. However, the researchers who employ these methods for entirely appropriate uses also tend to be the same people who use molecular phylogenies for taxonomic purposes, and their enthusiasm for these statistical methods has spilled over into taxonomy. But there is an important difference in these applications of phylogenetics that is often missed; classifications cannot incorporate phylogenetic uncertainty, and therefore an extra step in the process is required—assessment of conflict among different characters. Unfortunately, the recent rapid advances in statistical molecular phylogenetics have not been accompanied by similar advances in methods for exploring character conflict in molecular data.

Molecular phylogenetics usually proceeds in a stepwise fashion from data collection to sequence alignment to phylogenetic inference. At this point it is typical to assess the degree of support for the tree. For distance-based and optimisation methods this usually consists of bootstrap or jacknife resampling. For Bayesian methods, posterior probabilities generally serve this purpose. What is far less common is to examine whether there are any conflicting signals in the data. An optimal tree contains the relationships that have the greatest support, along with those other relationships that also have some support and are compatible with those that dominate. However, it is entirely possible that there are other signals in the data that are not represented in the optimal tree as they are not compatible with signals that have greater support. Relationships that are compatible with those that have the most support will appear in the optimal tree, and hence in classifications that are derived from it. Relationships that are not compatible will not be represented in these classifications, even if they have more support than some of the relationships that are compatible. It is therefore absence of conflict, rather than strength of support, that determines whether a particular relationship appears in a classification derived from an optimal tree. Lack of resolution in a tree may be due to lack of signal, or to a balance between conflicting signals of more or less equal strength. It is rare for molecular phylogeneticists to investigate which of these is the case, even when they intend to use the phylogeny for taxonomic purposes. Even for strongly supported relationships, there may be conflicting incompatible relationships with almost as much evidence in favour of them, but these would not appear in the optimal tree. By uncritically accepting the optimal tree without any further exploration of the data, these conflicting signals will be missed. It is important to investigate these conflicting signals as there may be very good reasons for supposing that the characters that support them are more likely to represent the true evolutionary history of the group in question, because they are more likely to reflect homology. This kind of data exploration is rare in molecular taxonomy, despite being the norm in morphological taxonomy, and this is one of the principal complaints of Mooi & Gill. This deficiency may be in part due to the enormous number of characters that molecular taxonomists have to deal with. This is recognised by Mooi & Gill who say that they "are aware that molecular workers feel that their work provides far too many synapomorphies to show node by node" (p.37) but then go on to criticise these workers for seeing "a request... to show evidence in support of their hypotheses... as unreasonable" (p.37). Mooi & Gill make a good point, but with the recent surge in the use of whole genome sequences for phylogenetics (Delsuc 2005, Philippe 2005) perhaps they underestimate the seriousness of this problem. It is indeed true that scientists should present evidence in support of their hypotheses, but when the number of characters is in the millions, as is starting to be the case with phylogenomic data sets, there is an urgent need for new ways of presenting this information.

Another potential reason for the lack of exploration of character conflict in molecular data is likely to be a perceived dearth of methods for making this job more manageable, and a misperception that the statistical methods that are so useful in other circumstances can be used to equal effect in a taxonomic context. In fact, methods do exist for exploring conflict in molecular data, but these are not widely used, and seldom applied in the context of molecular taxonomy. There is a serious need for molecular taxonomists to take up these existing methods and to put as much effort into developing new tools for this purpose as has been put into developing statistical methods for phylogenetic inference in other contexts.

Morphological taxonomists already have at their disposal a suite of sophisticated methods for exploring conflict in their data, and there is no reason why these cannot also be extended to molecular data. There is a rich literature in this area and therefore I will not attempt to summarise it here, but I will instead briefly describe four other approaches that may be particularly applicable to exploring conflict in molecular data, but which may be unfamiliar to some taxonomists, particularly those with little background in molecular biology.

Four tools for exploring character conflict in molecular data

1. Spectral analysis

Spectral analysis (Hendy & Penny 1993) is a graphical tool for visualising the phylogenetic signal in a particular data set. The strength of support for each split[1] is depicted as a bar graph, and splits are shown left to right in order of how strongly the data support that split. Crucially, the strength of support for relationships that conflict with the split is also given below the x-axis (Fig. 1). This means that a spectrum produced in this way contains visual information about the relative strength of support and conflict for each split. This makes spectral analysis an ideal tool for exploring character conflict in molecular data, but despite the fact that this method has been around for a long time, it has remained largely neglected. A rare example of its use in fish systematics is Lockhart et al. (1995). Spectral analysis is implemented in the computer program Spectrum (Charleston 1998).

2. Phylogenetic networks

Phylogenetic networks are a class of methods for reconstructing the evolutionary relationships of taxa that cannot be represented as a non-reticulating tree, e.g. due to hybridisation or horizontal gene transfer, etc. They can also be used as graphical tools for visualising phylogenetic uncertainty in data sets arising either from lack of phylogenetic signal or from character conflict. Networks have a wide variety of uses in systematics and evolutionary biology (Huson & Bryant 2006), but as uncertainty can arise from conflicting phylogenetic signals, network methods are particularly useful for exploring character conflict in molecular data. Rather than representing phylogenetic uncertainty in the form of incompletely resolved phylogenies containing polytomies, networks show unresolved relationships as parallelograms (Fig. 2). These parallelograms indicate parts of the tree for which there is conflicting phylogenetic signal. The lengths of the sides of the parallelograms indicate the relative strength of support for each of the alternative relationships. This approach has been more widely taken up than spectral analysis as a method for exploring character conflict (perhaps because it represents a development of existing tree-based methods of analysis, which systematists are already familiar with, rather than a whole new approach), but it tends to be restricted to studies where the primary goal is to understand the process of molecular evolution (e.g. Wang et al. 2010) rather than to establish phylogenetic relationships for taxonomic purposes. A recent ichthyological example is Marková et al. (2010). There are a number of different methods for constructing phylogenetic networks. Many of these are implemented in the computer program SplitsTree (Huson & Bryant 2006).

1. A split is defined as a branch that separates the taxa into two groups. For example, for five taxa, A-E, (A,B)(C,D,E) and (A,B,C)(D,E) represent compatible splits as they are both consistent with the tree ((A,B),C)(D,E), while (A,B)(C,D,E) and (A,C)(B,D,E) represent conflicting splits as there is no tree that can be simultaneously compatible both.

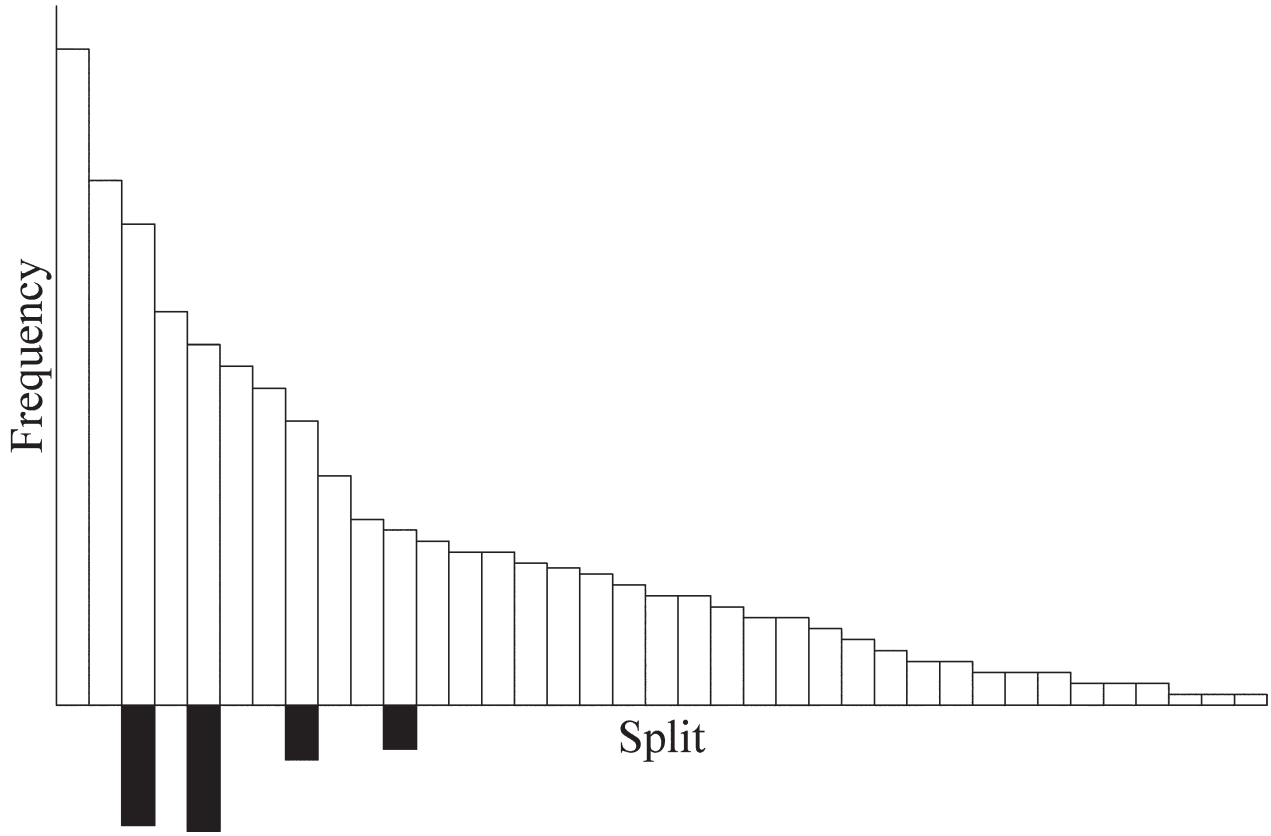

FIGURE 1. Spectral analysis results for a hypothetical data set. Each bar represents a different split in the tree. Bars above the x-axis represent the relative degree to which the data support that split. Bars below the x-axis represent the relative degree to which the data support relationships that conflict with (i.e. are incompatible with) that split. In this example, there is significant phylogenetic signal for relationships that conflict with splits 3, 5, 8 and 11.

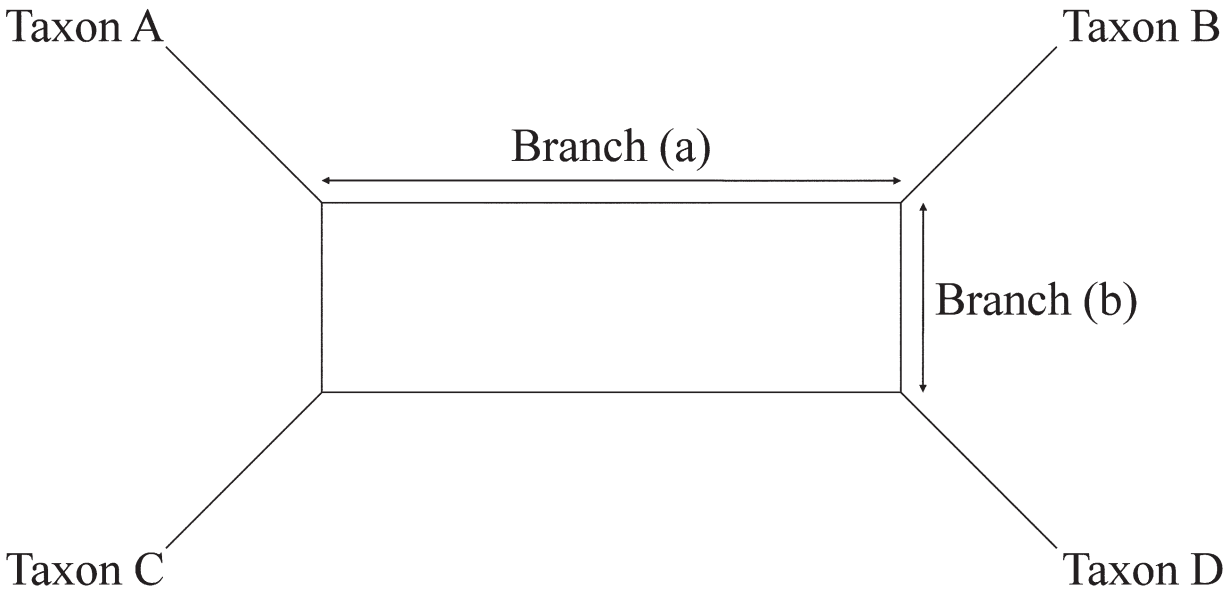

FIGURE 2. A phylogenetic network for a hypothetical data set. This network represents the relationships between four taxa, A-D. The length of branch (a) is proportional to the strength of support for the relationship (A,C)(B,D). The length of branch (b) is proportional to the strength of support for the relationship (A,B)(C,D). In this example there is conflicting support for both of these arrangements, but more weight is given to (A,C)(B,D) than to (A,B)(C,D).

3. Likelihood mapping

Likelihood mapping (Strimmer & von Haeseler 1997) is a graphical tool for visualising the phylogenetic signal in a data set of aligned sequences. For any four taxa, likelihood mapping represents the relative strength of support for the three possible fully resolved relationships between these taxa, i.e. (A,B)(C,D), (A,C)(B,D) or (A,D)(B,C), as a point on a triangle (Fig. 3). This method has mostly been used to assess whether a particular data set is suitable for phylogenetic analysis by plotting the distribution of points generated by every quartet of sequences in the data set across different regions of the triangle. If a significant proportion of points cluster in the corners then the data contain strong phylogenetic signal. If most points cluster in the middle then they do not.

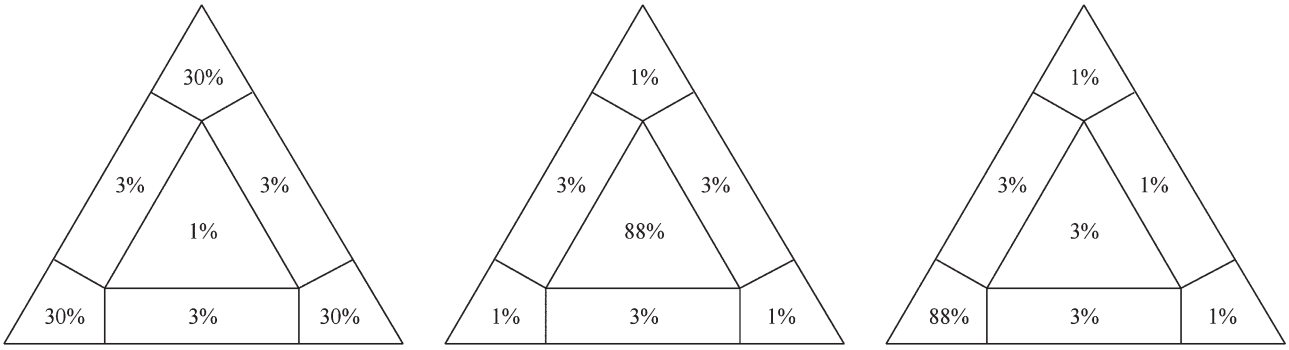

FIGURE 3. Likelihood mapping results for three hypothetical data sets. Each corner of the triangle represents a different phylogenetic hypothesis for the arrangement of four taxa. Figures represent the percentage of dots in each part of the triangle, where dots represent quartets of sequences sampled from an alignment of DNA sequences. In (a) most points cluster in all three corners; in the general approach to likelihood mapping this would represent a dataset with strong phylogenetic signal, in the context of testing a specific hypothesis it would represent a data set in which the results depend on which taxa are sampled from each group. In (b) most points cluster in the middle; in either approach this would represent a data set with little or no phylogenetic signal. In (c) most points cluster in one corner (lower left); this would represent a strong and consistent phylogenetic signal in favour of one particular hypothesis.

An alternative use of likelihood mapping is to define four particular groups of interest and then plot points for each quartet of sequences sampled from these groups. This allows an analysis of the support for a particular phylogenetic hypothesis as each corner represents one of the three possible arrangements of these groups. As before, if most points cluster in the middle then there is little phylogenetic signal. If most points cluster in one corner then there is a strong and consistent phylogenetic signal, but if points cluster in different corners then the results will depend on which taxa are sampled from each group. With careful modification (e.g. by analysing and comparing different subsets of the data) it should be possible to use likelihood mapping as a tool for exploring character conflict in molecular data, but this application of the method has not been widely explored. An example of the use of this approach in fish is Crow et al. (2006).

In fish systematics, likelihood mapping has mostly been used in a general way to assess the overall phylogenetic signal in particular datasets, rather than to investigate specific phylogenetic hypotheses. A recent ichthyological example is Sturmbauer et al. (2010). Likelihood mapping is implemented in the computer program TREE-PUZZLE (Schmidt et al. 2002).

4. Sliding window analyses

An important aspect of character conflict within genes is the spatial distribution of the conflicting characters. For sequences that result from recombination, conflicting signals may be spatially partitioned at different ends of the gene. Sections of a gene with different functions (e.g. parts of genes for trans-membrane proteins corresponding to hydrophobic and hydrophilic regions of the molecule) may also differ with respect to their phylogenetic signal if, for example, some parts are under strong selection pressure resulting in spurious signal reflecting convergent evolution of a common function rather than common ancestry. Methods are available to facilitate the spatial analysis of phylogenetic signal within a gene. In particular, sliding window analyses can be useful for identifying gene regions

with conflicting phylogenetic signals. At a larger scale this approach can also be used to investigate the spatial distribution of phylogenetic signal along genes on a chromosome or around a circular genome. Tools exist for this purpose (e.g. Proutski & Holmes 1998, Cai et al. 2005, Paraskevis et al. 2005), but it is usually a relatively simple matter to write scripts to automate this process for the particular phylogenetic software being used. These techniques have been designed mostly with analysis of recombination or natural selection in mind, but they are also applicable to the exploration of character conflict in phylogenetic analysis, although this potential use of these methods has not been extensively pursued in fish.

Previous phylogenies as priors in Bayesian analyses

Another, rather different, criticism made by Mooi & Gill of the way that molecular phylogenetic studies are usually carried out is that "statistical programs and optimisation have become the new authority figures, and when presented with two conflicting phylogenies the one including more 'data'... almost always the most recent – becomes the preferred topology regardless of previous evidence. This has led to an approach where phylogenies of yesterday are left with nothing to contribute to phylogenies of today" (p.27). In fact, Bayesian methods provide a potential mechanism for the incorporation of previous phylogenies as priors in the analysis. The question then becomes, do the new data provide enough evidence to change our opinion from our prior expectation of the topology based on previous phylogenies of the group? It is usual in phylogenetic analyses to use uninformative priors. (This could be seen as an unwillingness of researchers educated in a frequentist framework to completely embrace Bayesian philosophy.) By doing so, each analysis starts from the beginning and assumes that we do not already know anything about the relationships of the taxa in question, when in fact there has often been a lot of painstaking work put in to elucidating these relationships. Care must be taken to ensure that we do not include data in our new Bayesian analysis that have been used to construct the prior phylogeny, as this would bias the results and lead to circular reasoning, but it seems that careful incorporation of previous estimates of phylogenies as priors in the Bayesian analysis of new data could be a way of reconciling old and new phylogenies, and preserving the value in older data sets and analyses. This would have the benefit that as new data are collected, they would only alter our conclusions if the conflicting phylogenetic signal were sufficiently strong. As evidence accumulates, and we become more and more sure of the phylogeny of the group, it would take increasingly more convincing evidence to the contrary for us to change our opinion. This idea is encapsulated in the phrase famously attributed to the cosmologist Carl Sagan "extraordinary claims require extraordinary evidence"[1]. This would mean that our molecular classifications would become increasingly stable, in contrast to the current situation as presented by Mooi & Gill of phylogenies "flip-flopping" haphazardly from one topology to another as different data sets are analysed without reference to the results of previous studies. Despite the apparent potential benefits of this approach, it has not been seriously explored (but see Alfaro et al. (2005) for a discussion of some of these issues).

Conclusions

Perhaps part of the reason why the methods for exploring character conflict outlined here have been largely overlooked is that in many cases these approaches were developed for the purpose of investigating molecular evolution, rather than for systematics or taxonomy, and therefore they have been reported and discussed in journals and conferences that are unlikely to be of interest to taxonomists. As molecular data have become easier to obtain, taxonomists not familiar with the details of molecular evolution have been rightly keen to incorporate them into their analyses. However, just as the correct interpretation of morphological data requires a deep understanding of the structure and function of the organisms under investigation, so correct interpretation of molecular data requires a thorough grasp of the mechanisms of molecular evolution. Morphologists are fond of criticising molecular biologists for their ignorance of the basic biology of the organisms they are dealing with (e.g. Dunn 2003), but if they intend to incorporate molecular data into their analyses in a meaningful and critical way then taxonomists who have primarily been trained in the analysis of morphological data have an equal responsibility to understand the

1. Although he appears to have adapted this from a quote by the sociologist Marcello Truzzi, "extraordinary claims require extraordinary proof". In either case, this is a rather Bayesian point of view.

fundamental principles of molecular biology, and to familiarise themselves with the full range of methods available for analysing DNA, RNA, and protein sequences. Since we cannot all be experts on everything, the most productive way forward would appear to be collaboration between morphological taxonomists and molecular biologists, based on a mutual understanding of the promises and potential pitfalls of all sources of data. This seems to be the most likely route to achieving the ideal of Mooi & Gill that morphology should be "recognized as an equal partner with molecules in phylogenetic reconstruction" (p.37).

Acknowledgements

I would like to thank Marcelo Rodrigues de Carvalho and Matt Craig for inviting me to participate in this special issue, my colleagues in the Lincoln University Molecular Ecology Lab for useful discussion on these topics, and Lincoln University for financial support.

References

Alfaro, M.E., Smith, D.K., Xia, X. & Yuen, K. (2005) The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution and Systematics*, 37, 19–42.

Cai, J.J., Smith, D.K., Xia, X. & Yuen, K. (2005) MBEToolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution. *BMC Bioinformatics*, 6, 64.

Charleston, M.A. (1998) Spectrum: spectral analysis of phylogenetic data. *Bioinformatics*, 14, 98–99.

Crow, K.D., Stadler, P.F., Lynch, V.J., Amemiya, C. & Wagner, G.P. (2006) The "fish-specific" Hox cluster duplication is coincident with the origin of Teleosts. *Molecular Biology and Evolution*, 23, 121–136.

Delsuc, F., Brinkmann, H. & Philippe, H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6, 361–375.

Dunn, C.P. (2003) Keeping taxonomy based in morphology. *Trends in Ecology and Evolution*, 18, 270–271.

Edwards, S.V. (2009) Is a new general theory of molecular systematics emerging? *Evolution*, 63, 1–19.

Hendy, M.D. & Penny, D. (1993) Spectral analysis of phylogenetic data. *Journal of Classification*, 10, 5–24.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294, 2310–2314.

Huson, D.H. & Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23, 254–267.

Lockhart, P.J., Penny, D. & Meyer, A. (1995) Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. *Journal of Molecular Evolution*, 41, 666–674.

Marková, S., Šanda, R., Crivelli, A., Shumka, S., Wilson, I.F., Vukić, J., Berrebi, P. & Kotlík, P. (2010) Nuclear and mitochondrial DNA sequence data reveal the evolutionary history of *Barbus* (Cyprinidae) in the ancient lake systems of the Balkans. *Molecular Phylogenetics and Evolution*, 55, 488–500.

Mooi, R.D. & Gill, A.C. (2010) Phylogenies without synapomorphies – a crisis in fish systematics: time to show some character. *Zootaxa*, 2450, 26–40.

Paraskevis, D., Deforche, K., Lemey, P., Magiokinis, G., Hatzakis, A. & Vandamme, A.-M. (2005) SlidingBayes: exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. *Bioinformatics*, 21, 1274–1275.

Philippe, H., Delsuc, F., Brenner, H. & Lartillot, N. (2005) Phylogenomics. *Annual Review of Ecology, Evolution and Systematics*, 36, 541–562.

Proutski, V. & Holmes, E. (1998) SWAN: sliding window analysis of nucleotide sequence variability. *Bioinformatics*, 14, 467–468.

Schmidt, H.A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18, 502–504.

Strimmer, K. & von Haeseler, A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 94, 6815–6819.

Sturmbauer, C., Salzburger, W., Duftner, N., Schelly, R. & Koblmüller, S. (2010) Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data. *Molecular Phylogenetics and Evolution*, 57, 266–284.

Wang, D., Zhong, L., Wei, Q., Gan, X. & He, S. (2010) Evolution of MHC class I genes in two ancient fish, paddlefish (*Polyodon spathula*) and Chinese sturgeon (*Acipenser sinensis*). *FEBS Letters*, 584, 3331–3339.